The MINDS Workshops

- Two short workshops (November 12-13, 2007 and February 25-26, 2008) attended by around 25 people each time

- Sponsored by Heather McCallum-Bayliss of DTO

What is MINDS??

- Machine Translation
  - Discussion leader: Alon Lavie, Carnegie-Mellon U.
- Information retrieval
  - Discussion leader: Jamie Callan, Carnegie-Mellon U.
- Natural language processing
  - Discussion leader: Liz Liddy, Syracuse University
- Data resources
  - Discussion leader: Martha Palmer, U. of Colorado
- Speech understanding/transcription
  - Discussion leader: Janet Baker, Saras Institute/MIT

Overall goals

- Create a research agenda that is motivated ONLY by what each research area thinks is important to ITS goals, NOT by what they think would interest funders!
- Output of workshop
  - 10ish page report answering 2 questions
  - Report circulated to a wider group within the community for discussion
- One constraint: the research needs to be in the "aid" of information discovery

Questions

1) Make a list of 5-10 research discoveries that have led to a major paradigm change in your field.

2) Using this list as a guide, create a list of 5-10 research areas that would result in equally important paradigm shifts.

3) Additionally the second workshop looked at cross-area research by each pair of groups meeting together

What next

- DRAFT reports on:

  http://www.itl.nist.gov/iaui/894.02/minds.html

MINDS Workshops
MT Sub-group

Discussion Leader: Alon Lavie (CMU)

Other Members:

- David Yarowsky (JHU)
- Kevin Knight (ISI)
- Nizar Habash (Columbia)
- Chris Callison-Burch (Edinburgh)
- Teruko Mitamura (CMU)

The Big Paradigm Shift in MT

- **From** manually crafted rule-based systems with manually designed knowledge resources
- **To** search-based approaches founded on automatic extraction of translation models/units from data and language "features" that are extracted from vast amounts of online resources
- Some specific milestone "discoveries":
  - Sentence alignment for creation of parallel corpora
  - The "noisy-channel" model à IBM models of word alignment
  - Statistical Language models
  - Algorithms for extraction of phrase-to-phrase correspondences
- Several major developments enabled this shift:
  - Advent of enabling data and computational resources
  - Similar paradigmatic approaches in our sister fields – Speech and IR (inspired IBM models for SMT in early 1990s)
  - Advent of automatic MT evaluation metrics that support training and development

Current State-of-the-art in MT

- Search-based MT paradigm is well established:
  - SMT (phrase-based and now syntax-driven), EBMT, CBMT, Transfer, rule-based…
  - Common general framework:
    - Models for representing units of translation
    - "Decoder" that searches a large space of hypothesis combinations using a scoring function and selects a "final" translation
  - Different approaches to modeling and finding units of translation (TMs), learning or acquiring them from data, combining them together into complete hypotheses, and decoding.
  - Different representations.

So Why is MT Still so Bad?

- Combination of two fundamental problems:
  1. Weak Models: current Translation Models aren't strong enough to consistently generate correct translations
  2. Weak Discrimination: available knowledge resources are insufficient for effectively discriminating between good translations and bad translations
- Resulting Consequences:
  1. "Slim Pickings": The hypothesis spaces that are generated by current MT approaches often do not contain correct, or even good possible translations of the input
  2. "Finding the Needle in the Haystack": Our decoders aren't good enough to identify and select the good translations even when they are present in the search space

Major Research Priorities

- Objective #1: High Coverage MT for Many More Language Pairs:
  - Quality robustness across domains and genres within the same source language
  - Not just MT from Arabic and Chinese TO English:
    - MT from English
    - MT from and to low resource languages

Major Research Priorities

Some Proposed Technological Advances:

1. Better Translation Models
   - Research on specific sub-problems
2. Overcoming the resource acquisition bottleneck
   - Learning more from less data
3. More Discriminant Language Models
   - Beyond word-level ngrams
4. Multi-Engine MT
   - “One size” does not fit all

Fundamental Modeling Problem in MT

The “intermediate unit” problem:

- Parallel sentences are good translations of each other
- Word alignment algorithms can find word-level correspondences that are globally fairly reasonable
- But – translating complete “seen” sentences (Translation Memory) doesn’t generalize, and word-to-word translation doesn’t capture what’s truly necessary for MT
- Core challenge of finding good sub-sentential compositional units of translation and how they are composed is still not well understood: some pieces of meaning are compositional, others are not, and this differs from language to language…

Conclusion

Machine Translation:

- We can do it... You can help!
- Nosotros lata hacer ella , usted lata ayuda!
- Noi inscatolare fare lo , puoi aiuto!
- Nous can font le , vous can aider!
- Wir könnt ausführen es , Sie können abhelfen!
- Мы мочь делать он , ты мочь помогать!
- εμείs μπορώ κάνω αυτό , εσύ μπορείs βοήθεια!
- ويمكننا ان نفعل ذلك ، يمكنكم المساعدة!
- אנחנו יכולים להצליח, אתם יכולים לעזור!

The MINDS Workshops: Information Retrieval

Chantilly, VA

- **Jamie Callan (Chair)**
  - Carnegie Mellon Univ
- **James Allan**
  - Univ of Mass, Amherst
- **David Evans**
  - Clairvoyance Corp
- **ChengXiang Zhai**
  - Univ of Illinois, UC
- **Mark Sanderson**
  - Univ of Sheffield

Marina del Rey, CA

Jamie Callan (Chair)
  - Carnegie Mellon Univ
- **Charlie Clarke**
  - Univ of Waterloo
- **Susan Dumais**
  - Microsoft Research

Challenge 1:
Heterogeneous / Everyday Data

- **IR has mostly studied well-edited text**
  - E.g., most TREC corpora
  - Maybe still a good model of enterprise search
- **Many people’s information includes email, IM, social networks, blogs, pictures, videos, ….**
  - Heterogeneous across many characteristics
  - Eg., personal, trusted, noisy, adversarial
- **A very major change that probably requires…**
  - New retrieval models
  - New evaluation corpora and methodologies

Challenge 2:
Search Engines for HLT Apps

- **Many interesting NL applications draw information from large text corpora**
  - E.g., QA, MT, Speech, …
- **Today**
  - Bag of words search + HLT-oriented post-processing
  - Roll-your-own data structures and access methods
- **Tomorrow…**
  - Search engines that store text annotations & metadata
  - Query languages that support HLT access
  - Indexes that provide efficient access

Challenge 2:
Search Engines for HLT Apps

Examples

- **QA**
  - Use structured queries to match heavily annotated text
  - Ranked retrieval, <u>fast</u> retrieval
- **Speech**
  - Use a small language model to retrieve documents
  - E.g., to drive adaptive language models
- **MT**
  - N-gram frequency, completions, soft match

Search engines as “language databases”

Natural Language Processing

- **Eduard Hovy** – ISI, University of Southern California
- **Liz Liddy** – CNLP, Syracuse University
- **Jimmy Lin** – College of Information Studies, University of Maryland
- **John Prager** – IBM Research
- **Dragomir Radev** – School of Information, University of Michigan
- **Lucy Vanderwende** – Microsoft Research
- **Ralph Weischedel** – BBN

1. Machine Reading

- **Challenge:** Although most of the world's knowledge is available in text resources,
  - Software today cannot improve its effectiveness on a task through reading and learning from those texts
- **Today:** Software experts & knowledge engineers meticulously, manually improve system performance by adding knowledge
- **Future:** Robust NLP + Machine Learning offers the potential to bridge gap from text à knowledge, but need to be able to learn:
  - Encyclopedic lexical knowledge
  - Domain & genre structure
  - Mapping between language and knowledge representation

2. Socially-Aware Language Understanding

- **Challenge**: Incorporate social-context understanding in a system's interpretation of language
  - Requires system to accomplish deeper levels of interpretation
    - Discourse & Pragmatics
  - Beyond literal meaning – connotative as well as denotative
    - Politeness, sarcasm, humor, etc
- **Future**: Personalized NLP – Conversational systems that self-adapt to the person & the context
  - Ability of agent & person or 2 agents to jointly construct meaning
    - Each having own experience and expertise, but ability to take other's perspective into account to understand
    - Use of subtle features that highlight human-like linguistic intuitions to better understand and communicate

3. Annotation Science

- **Challenge:** Every HLT application for which rich training data is increased à performance improves
  - Need scientific basis / methodology for deciding:
    - What it is we need to annotate
    - Appropriate representation for the annotation
    - How the annotation will best be accomplished
  - Requires capability of mapping from 1 representation to another
    - Establish an interchange standard / an interlingua

- **Future:** A range of creative ways to acquire annotated training data:
  - Leveraging human capital on the Web
    - Social tagging – ESP tagging / Open Minds
  - Active learning as a methodology
  - Performance improved as depth & breadth of annotation builds

MINDS Workshop Data Resources:

Transcribed speech, Hansards, Treebank, WordNet, TREC corpora

Martha Palmer, U of Colorado
Stephanie Strassel, LDC, UPenn
Randee Tangi, Princeton

Are we done?

Ether a go-go (EAG) K(+) channels have been shown to be involved in tumor generation and malignant growth.

(PMID: 15364405)

Best NE F-measure: 83% (Dingare et al. 2005)

#1 - Science of Annotation and #2 - Annotation Infrastructure

- Methodologies and Best practices for
  - Choosing Corpora, Determining Annotations, Training Annotators, Evaluating results, …
- Portable, language independent, public domain annotation tools that produce standardized formats
  - Interoperable formats
  - Principles for layering annotations
- Community consensus on priorities

#3 - Closer integration of Emergent Technology Annotate SMARTER

- Machine learning desiderata for training data (negative examples?, contrast sets?,…)
- Immediate access to improved stochastic taggers for data pre-processing
- Dynamic access to active learning for isolating high payoff instances for annotation
  - Classifiers – currently in use for WSD
  - More complex tasks (syntactic parsing)????

Error Analysis!

DATA RESOURCES

MINDS Workshops-Speech Understanding/Transcription

Janet M. Baker, Li Deng, James Glass, Sanjeev Khudanpur, Chin-Hui Lee, and Nelson Morgan

3-5 Year Research Programs: 1-2 of 6

- Everyday Audio: Unknown/New environment, channels, speakers, content in test data
  - Acoustic/Speaking environment: Reverberation, noise, overlapped speech
  - Channel used for speech capture: far-field microphone, cellular phone
  - Speaker characteristics and speaking style: nonnative accent, emotional speech
  - Language characteristics: sublanguages and dialects, vocabulary, genre and topic
  - Links to brain & cognitive science, natural language processing, IR
- New language with limited annotated resources, possibly a lot of unannotated resources
  - Generalization (e.g., cross-language features, phone sets, lexicons), adaptation
    - Speech/Acoustic units that are more language-universal than phones
    - Cross-lingual language modeling
  - Links to machine translation, natural language processing, document understanding, IR

3-5 Year Research Programs: 3-4 of 6

- Unsupervised/semi-supervised language acquisition by the system
  - Pattern discovery, generalization, active learning, adaptation
  - Language acquisition from multi-sensory cues, and interaction with the environment
    - E.g. hearing a person/place name in speech, then discovering it in related text
  - Links to brain & cognitive science, natural language processing, information retrieval
- Recognize low frequency events
  - Rare/Unexpected events can be important to recognize – yet ignored by current metrics
  - Rare words often misrecognized as other similar words (the unknown unknown)
  - Correct recognition requires confidence, uncertainty modeling
  - Links to cognitive science, natural language processing, information retrieval

3-5 Year Research Programs: 5-6 of 6

- Gain insights into how the brain processes speech and language
  - information vs. signal processing
  - Focus on learning from scientific knowledge: adaptation rates to new environments, accented speech, role of episodic learning, attention … how do humans do it; how well do they do?
  - Leverage new developments in brain imaging, cortical processing of speech and language
  - Links to brain & cognitive science, natural language processing, information retrieval, document understanding
- Listening and writing comprehension tests (1st to 3rd grade)
  - Document and/or questions could be oral/written
  - Sentence segmentation, named entity extraction, partial information
  - Links to natural language processing, information retrieval, document understanding?

Cross-Research Areas: Meeting Room Scenario

- MT – Translate text to/from different languages
- IR – Data Mining, create LMs, etc.
- NLP – Create summaries, Tag named identities, etc.
- Data Resources – Collect & Tag Train/Test Mtls.
- Speech Understanding – Transcription, Metadata, etc.

Cross Research Area System Enablers

- Utilize *Multiple* Knowledge Sources
- Design *Flexible Parallel* System Architectures for Fault Tolerance and Robustness
- Create *Some Common* Cross Research Area Corpora and Tools
- Compare and Contrast *System* Evaluations with *Complexes* of *Cooperative* Components

Cross Research Areas: Application Opportunities

- Newsroom – Collection, Summarization, Editting and Production (Audio,Video,Text)
- Quarterly Investor Teleconferences
- TV/Video Closed Captioning
- Battlefield Command Headquarters
- Multinational Corporate Analysis, Planning and Operations